基于持续增量模型的低速端口扫描检测算法 *

沈　晶，薛少勃，刘海波†
(哈尔滨工程大学 计算机科学与技术学院, 哈尔滨 150001)

摘　要：端口扫描是一种常见的有效入侵技术，用于搜索易受攻击的 Internet 主机和端口。快速端口扫描的检测技术已经成熟，但是隐蔽的低速端口扫描检测效果有待提升。针对低速端口扫描进行了研究，根据低速扫描的时间持续性和特征分散性，提出了一种基于持续增量模型的低速端口扫描检测算法，结合条件熵对特征分布的评估达到检测目的。实验结果表明算法的检测率能达到 99.78%，且误报率为 7%。算法适用于多种复杂网络环境，且不需要网络先验知识，检测率对阈值的精确性要求低，能够有效检测到低速端口扫描行为。

关键词：低速端口扫描；持续增量模型；信息熵；异常检测
中图分类号：TP393.08　　doi: 10.19734/j.issn.1001-3695.2018.10.0730

Low-speed port scan detection algorithm based on continuous incremental model

Shen Jing, Xue Shaobo, Liu Haibo†
(*School of Computer Science & Technology*, *Harbin Engineering University*, *Harbin* 150001, *China*)

Abstract: Port scanning is a common and effective intrusion technique for searching vulnerable Internet hosts and ports. The detection technology of fast port scanning has matured, but the hidden low-speed port scanning detection effect needs to be improved. This paper studies low-speed port scanning. According to the time persistence and feature dispersion of low-speed scanning, a low-speed port scanning detection algorithm based on continuous incremental model is proposed. The conditional entropy is used to evaluate the feature distribution. The experimental results show that the detection rate of the algorithm can reach 99.78%, and the false positive rate is 7%. The algorithm is applicable to a variety of complex network environments, and does not require network prior knowledge. The detection rate has low accuracy on the threshold, and can effectively detect low-speed port scanning behavior.

Key words: low-speed port scanning; continuous incremental model; information entropy; anomaly detection

0　引言

端口扫描是网络入侵的重要组成部分，端口的开放状态意味着通信渠道是否顺畅，攻击者通过发送试探性报文来发现目标系统存在的漏洞，并利用这些漏洞对目标主机进行攻击[1]。因此，有效的端口扫描行为检测可以将部分入侵行为扼杀在萌芽状态，从而达到防患于未然的效果，减少恶意攻击行为所带来的损失。

现在有众多的安全工具可以实现扫描一个范围的端口和 IP 地址。不过，一个入侵监测系统(IDS)一般将能够捕获这种明显的扫描行为,然后可以通过阻挡源 IP 地址来实现关闭这个扫描，或者自动向安全管理员告警。但是多数认真的攻击者一般不会通过执行这种扫描来暴露自己的意图；相反，他们会放慢速度，使用半连接(half-connection)尝试来找出你的可用资源[2]。尽管这种低速的攻击方法很耗时，它实现起来却不困难，更重要的是很难防范它。

对于快速的端口扫描多采用基于阈值的检测算法，snort 等大多数的入侵检测系统根据所配置的阈值信息(一定的时间段内允许某源地址所访问目的主机数和端口数的最大值)，实时统计网络主机所访问的主机数和端口数，如果超过所设定的阈值则为扫描行为。但是阈值容易受到环境影响，文献[3]提出了改进的状态阈值随机游走算法，利用网络服务的主动映射来考虑连接尝试失败的良性原因，能有效应对环境变化的干扰，显著降低误报率。文献[4]开发了一个基于规则的网络入侵检测系统，可在 snort 平台上使用。文献[5]提出了一种虚拟网络功能架构,用于在网络功能虚拟化云开放平台中检测端口扫描行为，对分布式端口扫描行为具有良好的检测效果。以上方法对于速度快的扫描行为可以实现很好的检测，但是难以检测低速的端口扫描行为。文献[6]采用模糊技术方法进行异常检测，具有一定的效果，但存在参数敏感问题。邵国林等人在文献[7]中提出一种基于 Dempster-Shafer 证据理论的检测方法，不需要训练精确的阈值，且能够检测不同速度的端口扫描攻击。文献[8]提出一种协调扫描检测算法，可对分布式扫描攻击作出有效检测。文献[9]用主成分分析方法量化端口扫描的风险指数，来检测低速端口扫描攻击。随机阈值算法是用于检测扫描仪的高效且广泛引用的方法，但是可以通过将探测尝试与已知活动主机的访问混合来规避它们。文献[10]提出一种与随机阈值互补的方法应对规避策略，通过目的主机的入度情况对源主机建立风险评估，结合源主机的出度评判攻击的可能性。但是该方法建立在目的主机存在活跃度差异的基础上，对于活跃性均衡的网络环境无法适应。

本文根据低速扫描的时间持续性和特征分散性，提出一种基于持续增量模型的低速端口扫描检测算法。算法选取两

收稿日期：2018-10-14；**修回日期：**2018-11-29　**基金项目：**国家重点研发计划资助项目（2017YFC0820700）；黑龙江省自然科学基金资助项目（F2018011）；中央高校基本科研业务费专项资金资助项目（HEUCFP201808）

作者简介：沈晶（1969-），女，黑龙江省鸡西人，副教授，博士，主要研究方向为机器学习、信息安全；薛少勃（1992-），男，硕士研究生，主要研究方向为信息与系统安全；刘海波（1976-），男（通信作者），副教授，博士，主要研究方向为智能计算与安全（liuhaibo@hrbeu.edu.cn）.

个有效特征组成一个二项集，通过观察二项集的持续性和多项集对应特征的增长情况细粒度的筛选出网络中的可疑流量。最后通过条件熵对特征分布的评估，检测低速端口扫描攻击。

1　网络流量特征选取

1.1　低速端口扫描行为描述

端口扫描通常分为水平扫描、垂直扫描和块扫描三种情况[11]。水平扫描是对多个不同目的主机的同一个端口进行扫描，垂直扫描是对特定目的主机的多个端口进行扫描，块扫描是水平扫描和垂直扫描的结合，是对多个不同目的主机的多个端口进行扫描。

有很多端口扫描工具，用户可以根据自己的需要选择不同的模板，由 Nmap 负责选择实际的时间值[12]。模板也会针对其他的优化控制选项进行速度微调。参数-T 有[0,5]的可选值，本文的数据集中出现以下三种：

-T 1：数据包的发送间隔是 1 5s。

-T 2：不增加太大的网络负载，串行每个探测，并使每个探测间隔 0.4 s。

-T 3：Nmap 的默认选项，在不使网络过载或者主机/端口丢失的情况下尽可能快速地扫描。

参数-T 3 对应的扫描攻击为常见的快速扫描，-T 1、-T 2 对应的攻击称为低速端口扫描。扫描的速度越小，攻击流量在单位时间内所占比重越低，检测的难度也就越大。本文的数据集中单位时间流量的数量平均值为 3 554 条每分钟，参数为-T 3 时攻击数据所占比例高达 95%。若攻击按照间隔 0.4 s 进行，则攻击出现的频率约为 0.04，如果间隔为 15 s，频率将更低。因此低速扫描攻击的隐蔽性更强，检测难度更大。

1.2　低速端口扫描特征分析

扫描攻击发生时源主机和目的主机的连接示例如图 1 所示。其中 S 表示源主机，D 表示目的主机，C 表示与该源主机建立连接的不同目的主机的数量。正常活动的源主机通常情况下符合 S1 和 S2 的行为模式，特点是这类源主机和不同目的主机间建立的连接数少；而攻击主机更符合 S3 的情况，其和不同的目的主机建立大量的连接。

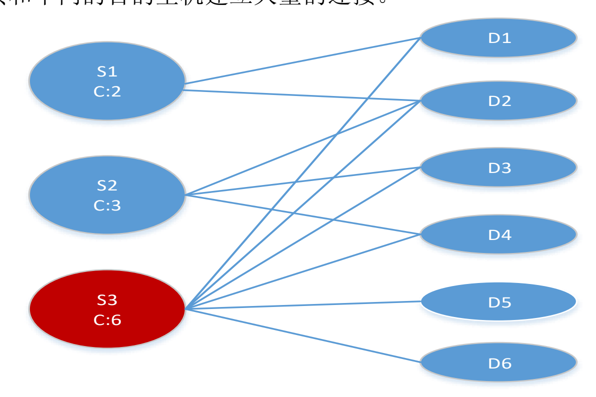


图 1　端口扫描攻击连接示例

Fig. 1　Port scan attack connection example

当然不能仅从连接数的大小来区分攻击和非攻击。例如服务器要响应大量请求时就会反向对请求者发送报文，这时的服务器主机的 C 值会很大，但服务器并没有发动攻击，因此 C 值大的源主机不一定是攻击者。对攻击行为的判定需要有更多依据。

本文用 sip 表示源主机 IP 地址，dip 表示目的主机 IP 地址，dpt 表示目的端口。攻击发生时相同 sip 和 dip 的报文采用的协议和报文长度具有局部一致性，在一定时间内报文的协议类型通常相同，多为 TCP，UDP 和 ICMP。由于端口扫描的主要用于侦测端口的开放状态，只需要少量的报头信息就能完成。为了减小网络开销攻击者，通常将报文的数据部分控制在较低范围，报文长度多在 300 Bytes 以下。另外低速扫描行为还有两个明显的特性：

a)持续性。由于采用的是低速的隐蔽形式来扫描端口，这种攻击就要以时间为代价，所以攻击具有持续性，水平扫描的 sip-dpt 组合和垂直扫描的 sip-dip 组合持续时间通常超过 10 min。

b)分散性。水平扫描的 sip-dpt 组合对应的 dpt 总量大，整体上呈现松散分布，重复率低；垂直扫描的 sip-dip 组合对应的 dip 总量大，整体上呈现松散分布，重复率低。

正常流量的分布情况与低速扫描行为在持续性和分散性上有明显的区别，合理利用这两个特征能够有效筛选出可疑流量，为异常检测的快速进行提供帮助。

2　异常流量检测算法

本文首先用单一特征信息熵方法，挖掘出明显的异常流量；然后重点针对低速扫描攻击作出检测。根据端口扫描特点的分析，将攻击分为水平扫描和垂直扫描，用持续增量模型过滤使检测范围缩减到最小，再用二项集组合的条件熵作最终的判定。

2.1　信息熵检测方法

香农在 1948 年将热力学中熵的概念引入到信息论中，定义为信息熵[13]。信息熵用来解决信息度量的问题，公式如下：

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \tag{1}$$

其中：$H(X)$ 代表了随机变量 X 的信息熵；$p(x_i)$ 代表随机事件 x_i 出现的概率。网络流量数据由离散信息源组成，熵可以有效度量系统参数分布的变化情况，描述长时间的随机过程以及网络流量在某些维度上的分布状况。基于熵的异常检测系统的主要思想是:一旦有异常流量发生，总体流量的熵值会随之发生变化，通过熵值的变化检测出该异常。

定理 1　切比雪夫不等式：设随机变量 X，$E(X)=\mu$，$D(X)=\sigma^2$，则对任意的 $\varepsilon>0$，必有

$$P(|X-\mu|\ge\varepsilon)\le\frac{\sigma^2}{\varepsilon^2} \tag{2}$$

将流量的信息熵看做随机变量，根据切比雪夫不等式，令 $\varepsilon=2\sigma$，则有 $P(|X-\mu|\ge 2\sigma)\le 1/4$。当攻击发生时，特征的熵值往往偏向（0,1）的两端。如果将阈值设置为 $\mu\pm2\sigma$，则异常流量落在这个区域的概率小于 25%。

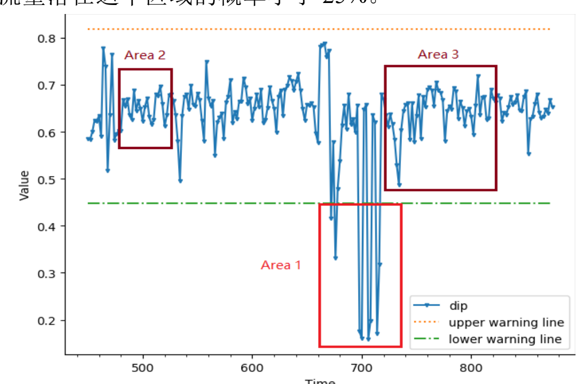


图 2　特征 dip 信息熵折线图

Fig. 2　Information entropy line graph of feature dip

图 2 为特征 dip 的信息熵折线图。图中横坐标表示时间，单位为 min；纵坐标表示特征 dip 在各个时间窗口内的熵值，

经过归一化运算其范围在(0,1)之间。图中 upper warning line、lower waring line 的纵坐标分别代表信息熵上、下两个阈值，落在两条线之间的坐标点被判定为正常点，两条线外的点为可疑点，阈值由各个时间窗口特征的信息熵根据式(2)计算得到。方框 Area1 圈中的坐标点熵值低于设定的阈值，被判定为异常流量。出现较大波动的原因可能是当前时间在午休范围内，网络波动大符合人们的行为规律，但也可能确实发生了攻击。通过频繁项集支持度计数的方法能够分辨出其中部分流量符合 DOS 攻击行为。Area2 和 Area3 中的坐标点波动幅度与 Area1 相比，根据阈值信息将其判定为正常流量，但真实情况是这两个区域中存在水平扫描类型的低速扫描攻击。由此可见单一特征的信息熵检测方法无法检测到低速扫描攻击，需要找到更精准的辨别方法。

2.2　持续增量模型筛选方法

单纯用某些独立特征的信息熵来做检测，不考虑特征间的联系，只能检测到熵值变化很大的攻击。提升检测率的有效途径是挖掘流量中潜在的信息，特征间的关系就是一种重要的潜在信息。按照低速端口扫描的特征分析，本文从持续性和分散性入手建立关系模型。

对于水平扫描攻击，将数据集按时间顺序分为多个窗口，在每个窗口内选取 sip-dpt 特征字段作为一个二项集组合，统计 sip-dpt 组合对应的 dip 特征字段的分布情况。低速水平扫描攻击的 sip-dpt 组合具有持续性，sip-dpt 对应的 dip 在数量上持续增长，直到完成一次扫描。而大部分正常流量的 sip-dpt 组合连接时间小于 1 min，不具有持续性，且正常流量的 sip-dpt 组合对应的报文数量和报文长度分布无规则，在每个窗口中 dip 的重复率高，dip 数量也很难持续增长。

低速扫描通过降低攻击速度隐藏自己，攻击发生时单位时间内攻击流量所占比重多低于 5%，容易被正常流量稀释。如果能将攻击数据的比重放大，减小正常流量的影响，那么异常检测的难度能降低。由于正常流量和低速端口扫描性质上的差异，如果按照持续性和分散性的标准过滤数据，那么过滤后的数据中正常流量所占比重会迅速下降。

按照上述思想，本文提出基于持续增量模型的低速端口扫描检测算法，用于考察流量的持续性和分散性。考察 sip-dpt 持续性的步骤如下：

a)以 2 min 为窗口长度分割数据。

b)统计每个窗口中不同的 sip-dpt 组合，并将对应的 dip 组成一个集合。

c)筛选出累计窗口数超过 5 次的 sip-dpt 组合。

找出持续 sip-dpt 组合后就要检验其对应的 dip 是否具有分散性。用增量过滤算法来考察分散性，对某个窗口数为 N 的持续组合，记时间窗口 i $(i<N)$ 内的 dip 集合为 dip_i。第 i 个窗口之前的 dip 集合的并集记为 u_dip。若当前窗口对应的 dip 集合中不存在与 u_dip 不同的元素，则 u_dip 清空。具体步骤如下：

a)初始化窗口号 $i=0$,记 dip_0 为 pre_dip，pre_dip 集合长度记为 len_0。

b) i 加 1，令 u_dip 为 pre_dip 与 dip_i 的并集。

c)求 u_dip 与 pre_dip 长度的差值 d，若 $d=0$，则 $pre_dip=dip_i$；否则 $pre_dip=u_dip$。

d)令窗口 i 对应的值 len_i 为 pre_dip 的长度。若 $i>=N$,结束；否则转到步骤 b)。

增量过滤算法将持续 sip-dpt 组合所对应的 dip 数量递增的流量筛选出来。持续增量模型筛选的结果如图 3 所示。每个子图的标题是一个 sip-dpt 组合。横坐标表示时间，纵坐标表示 sip-dpt 组合对应的 dip 的局部累加和 len_i。图中带有“*”标志的折线代表的是持续增量模型筛选的结果。折线图呈现周期性增长的态势，这是由于扫描分为多个阶段进行，一次扫描完成时 dip 数量达到峰值不再增加，这时需要重新开始新的计数周期用以探测下一次扫描。如果一个周期内的窗口数小于 5，那么这段窗口就不满足攻击的持续性，就可以排除掉，带有“o”标志的折线正是在原有折线图基础上剔除局部增长窗口数少于 5 的周期后得到的。

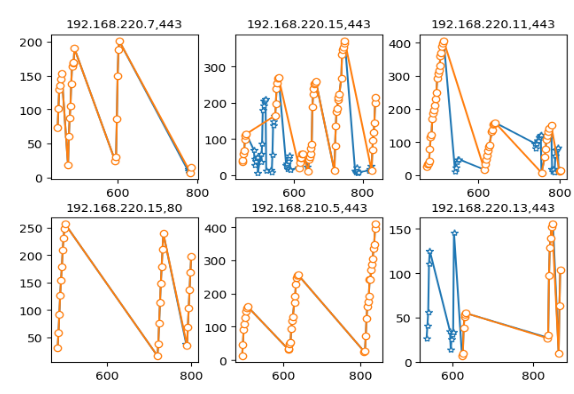


图 3　持续增量模型筛选结果

Fig. 3　Screening results of continuous incremental model

符合持续增量模型的组合被标记为可疑流量。最后用条件熵对每个增长周期做判定。正常流量的 sip-dpt 条件特征字段对应的 dip 分布集中，即使有小部分正常流量满足了持续增量模型，在局部增长区域内其 dip 的重复率通常也比低速端口扫描的高，因此熵值偏小。而水平扫描攻击针对的是指定端口的不同 dip，分布松散。因此可以将 dip 条件熵值的大小作为判断依据。候选组合的条件熵如表 1 所示。从表 1 中能够看到，攻击组合“192.168.220.15,80”对应的条件熵明显大于其他正常组合。当确定攻击源 IP 和受害者端口后，根据协议类型和报文长度很容易找到受害者 IP 地址。

表 1　候选组合的条件熵

Table 1　Conditional entropy of candidate combinations

sip-dpt 组合	dip 熵值	sip-dpt 组合	dip 熵值
192.168.220.7,443	0.629369	192.168.220.15,80	0.920444
192.168.220.15,443	0.570306	192.168.210.5,443	0.600042
192.168.220.11,443	0.58866	192.168.220.13,443	0.635613

对于垂直扫描攻击的检测方法和水平扫描类似，只需要调整特征字段的组合为 sip-dip，用持续增量模型筛选对应的 dpt，最后用条件熵确定攻击流量。

3　实验结果及分析

3.1　网络流量数据集

本文的数据来自 CIDDS(coburg intrusion detection data sets)[14]。CIDDS 是为基于异常的网络入侵检测系统创建评估数据集。入侵检测系统的发展可以看做是攻击者企图和防御者触发调整之间不断演变的过程。从这个角度来看，用旧数据集测试当前的入侵检测系统是不合适的。因此，CIDDS 的主要目标是生成可定制和最新的数据集。为了实现这一目标，CIDDS 背后的基本思想是使用 OpenStack 在虚拟环境中创建标记的基于流的数据集。

CIDDS 数据集由模拟的一个小型的小型商业环境产生，网络拓扑结构如图 4 所示。此环境包括多个客户端和典型服务器，如电子邮件服务器或 Web 服务器。Python 脚本用于模拟良好的用户行为，如浏览 Web、发送和接收电子邮件以及交换文件等。为确保尽可能真实的用户行为，使用配置文件

设置每个用户的特征。在网络中执行拒绝服务（DoS），暴力攻击和端口扫描来生成恶意流量。本文针对低速端口扫描攻击进行检测。

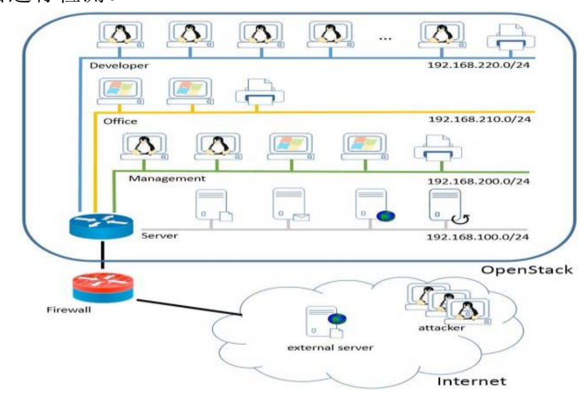


图 4 模拟环境拓扑图

Fig. 4 Topology diagram of simulated environment

3.2 实验结果及分析

由于 CIDDS 数据集总量大，本文只选取第二周的第二天 7:30 开始到下午 14:35 的 150 万条数据，数据包含了参数为 -T 1、-T 2 的低速水平扫描和垂直扫描攻击及 DOS 攻击，其中低速扫描攻击 15 096 条，占比约为 1%。本文采用检测率和误报率两个指标来评价模型的准确度。检测率是指检测到的端口扫描总数与实际端口扫描总数的比率；误报率是指将正常行为判断为端口扫描的总数与检测结果总数的比率。本文将基于持续增量模型的方法和基于端口比的方法进行了比较，实验结果如表 2 所示。

表 2 端口扫描检测方法对比

Table 2 Comparison of port scan detection methods

检测方法		阈值				
		0.65	0.70	0.75	0.80	0.85
持续增量模型	检测率	0.9978	0.9978	0.9978	0.9492	0.8742
	误报率	0.2074	0.1705	0.0700	0.0157	0.0109
端口比模型	检测率	0.9451	0.9431	0.9431	0.9418	0.8828
	误报率	0.3542	0.3163	0.2663	0.1980	0.1451

常见的基于端口比阈值的方法，通过一定的时间段内源地址所访问目的主机数和端口数的比值来判定是否是攻击。低速扫描虽然在攻击发生后总的扫描数量依然庞大，攻击数据的端口比通常高于 0.65，但是由于低速攻击持续时间长，正常流量和攻击流量混杂在一起无法作出有效区分，这种粗糙的判定方式容易造成正常流量的误判。从表 2 中能够看出端口比方法的检测率最高达到 94%以上，对应的误报率大于 19.8%，可见端口比方法对低速扫描攻击的检测效果不是很理想。

相较于端口比方法，持续增量模型的检测率提升了约 5%，而误报率下降了约 10%。当阈值在[0.65,0.75]区间上时，持续增量模型的检测率一直维持在 99.78%，由此可见持续增量模型的检测率对阈值的精确度要求较低。在误报率方面，随着阈值的增加，误报率逐渐降低，在保持检测率为 99.78%的条件下误报率能降至 7%。

持续增量模型充分利用特征之间的内在联系，构造二项集与对应特征的关系模型，精准刻画出低速端口扫描的持续性和分散性。正常流量的连接时间通常比低速扫描短暂，且通信端口分布在一定的范围内，访问的目的 IP 也有限，只有极少数的正常流量符合持续增量模型，经过筛选大部分正常流量被过滤出去,因此检测的范围将迅速缩小，检测难度也极大的降低。实验中过滤掉的正常流量占比超过 98.6%，说明了持续增量模型的高效性。持续增量模型在遇到特征的数量不再增加时会开始新的周期，这个机制使得异常检测分阶段进行，利用条件熵对特征分布情况的计算，能够精确区分一个连接中的攻击部分和非攻击部分。这些性质使得持续增量模型的检测率和误报率均优于端口比方法。

4 结束语

本文首先用单特征信息熵做粗分类，挖掘出明显的异常流量；然后重点针对低速扫描攻击作出检测。根据端口扫描特点的分析，将攻击分为水平扫描和垂直扫描，用持续增量模型过滤使检测范围缩减到最小，再用二项集组合的条件熵作最终的判定，分类后的检测结果更细致。采用持续增量模型能从图像上直观地看出攻击发生时扫描的强度，检测过程不需要先验知识，且检测率不受阈值精确度影响。

参考文献：

[1] 王龙业，罗杰. 互联网端口扫描攻击的安全检测方法 [J]. 信息安全与技术, 2016, 7 (2): 44-45, 64. (Wang Longye, Luo Jie. The overview of safety testing methods aimed at the internet port scanning attack [J]. Information Security and Technology, 2016, 7 (2): 44-45, 64.)

[2] 程巍，章磊，高传善. 隐蔽端口扫描的原理及防御方法 [J]. 计算机应用与软件, 2004 (8): 97-99. (Cheng Wei, Zhang Lei, Gao Chuanshan. The principle and defense of stealthy port scanning [J]. Computer Applications and Software, 2004 (8): 97-99.)

[3] Alsaleh M, Oorschot P C V. Revisiting network scanning detection using sequential hypothesis testing [J]. Security & Communication Networks, 2012, 5 (12): 1337-1350.

[4] Patel S K, Sonker A. Rule-based network intrusion detection system for port scanning with efficient port scan detection rules using snort [J]. Future Generation Communication and Networking, 2016 , 9(6)：339-350.

[5] Sanz I J, Lopez M A, Mattos D M F, *et al.* A cooperation-aware virtual network function for proactive detection of distributed port scanning [C]//Proc of Cyber Security in Networking Conference. Piscataway, NJ: IEEE Press, 2017.

[6] 吉治钢，蔡利栋. 一种基于端口扫描的模糊检测策略 [J]. 计算机应用, 2003 (10): 87-88, 92. (Ji Zhigang, Cai Lidong. A port-scanning detection method based on fuzzy set theory [J]. Computer Applications, 2003 (10): 87-88, 92.)

[7] Shao G, Chen X, Yin X, *et al.* A fuzzy detection approach toward different speed port scan attacks based on Dempster-Shafer evidence theory [J]. Security and Communication Networks, 2016, 9 (15): 2627-2640.

[8] Lv Y, Li Y, Tu S, *et al.* Coordinated scan detection algorithm based on the global characteristics of time sequence [C]// Proc of International Conference on Broadband&Wireless Computing. Piscataway, NJ: IEEE Press, 2018: 199-206.

[9] Park S, Kim J. A study on risk index to analyze the impact of port scan and to detect slow port scan in network intrusion detection [J]. Advanced Science Letters, 2017, 23 (10): 10329-10336.

[10] Harang R E, Mell P. Evasion-resistant network scan detection [J]. Security Informatics, 2015, 4 (1): 1-10.

[11] Bhuyan M H, Bhattacharyya D K, Kalita J K. Surveying port scans and their detection methodologies [J]. The Computer Journal, 2011, 54 (10): 1565-1581 .

[12] Anandita S, Rosmansyah Y, Dabarsyah B, *et al.* Implementation of

dendritic cell algorithm as an anomaly detection method for port scanning attack [C]// Proc of International Conference on Information Technology Systems and Innovation. Piscataway, NJ: IEEE Press, 2015.

[13] Bereziński, Przemysław, Jasiul B, *et al.* An entropy-based network anomaly detection method [J]. Entropy, 2015, 17 (4): 2367-2408.

[14] Ring M, Wunderlich S, Gruedl D, *et al.* Creation of flow-based data sets for intrusion detection [J]. Journal of Information Warfare, 2017, 16 (4): 40-53.